Project Atlas: Automated Market Segmentation A Hybrid Approach using Vector Space Clustering and OpenAI GPT-3

Abdullah Abid[1] and Yigit Ihlamur[2]

[1]University of Oxford, UK
[2]Vela Partners, US

March 2023

Abstract

The tool developed in this research project has the potential to revolutionize the way that companies are analyzed and understood. By automating the process of clustering and labeling similar companies based on their descriptions, the tool saves valuable time and resources for businesses and researchers. The use of word embeddings and hierarchical clustering allows for more accurate and nuanced grouping of companies, while the integration of natural language processing using OpenAI's GPT-3 language model provides insightful and descriptive labels for the resulting clusters. This tool could be applied to a variety of industries, such as finance, marketing, and consulting, where analyzing large sets of companies is crucial for decision-making. Furthermore, it has the potential to enable new types of research and insights by facilitating the exploration of previously unconsidered connections and patterns within the data. Ultimately, this tool represents a significant step forward in the field of company analysis, offering a more efficient and effective way to gain valuable insights into the landscape of businesses.

1 Motivation

The motivation behind this research project stems from the growing need for more efficient and effective methods of analyzing large sets of companies. Traditional methods of analyzing companies based on financial metrics can be time-consuming and often lack the nuance required to truly understand the differences and similarities between businesses. By developing a tool that leverages machine learning techniques and natural language processing, this project aims to provide a more streamlined and insightful approach to company analysis. Moreover, the potential applications of this tool in various industries make it an attractive prospect for businesses and researchers looking to gain a competitive edge in their respective fields. Ultimately, the motivation behind this research project is to provide a valuable resource that can help stakeholders better understand and navigate the complex landscape of companies in today's fast-paced business world.

2 Data Exploration

For this research project, the original data set consisted of one million rows of data, containing business names and IDs, as well as their corresponding business descriptions. However, to enable faster prototyping of the code, the data set was trimmed down to the first 5000 rows. This decision was made to reduce the amount of processing time and resources required to run the code on the entire data set, allowing for more efficient experimentation and development. Even with a smaller data set, the analysis yielded significant insights.

3 Methodology

3.1 Data pre-processing

Pre-processing data is a crucial step that involves cleaning, standardizing, and transforming the raw data into a format that can be effectively analyzed and modeled.

One of the essential steps in this process is the removal of common words, known as stop words, which do not carry much meaning and can potentially interfere with the analysis. In addition to removing stop words, a lemmatizer is often used to reduce words to their base form. This helps to reduce the complexity of the data and consolidate related words, which can improve the accuracy of the analysis.

By using a lemmatizer, variations of a word such as "run," "running," and "runner" are reduced to their root form "run." Furthermore, all punctuation is eliminated and all words are converted to lowercase to further standardize the data, ensuring that the analysis is not influenced by the style of the writing or the use of capitalization. Lower casing all words also helps to reduce the size of the vocabulary and simplify the input for the model, making it more efficient.

With the right pre-processing techniques, even the most unstructured and noisy text data can be transformed into meaningful and actionable insights.

3.2 Data Clustering

3.2.1 Embeddings

The pre-processed descriptions were fed into the Word2Vec neural network model to generate embeddings for each word in the descriptions. Word embeddings are dense vector representations of words that capture their semantic meaning and relationship with other words in a text corpus. The Word2Vec model was chosen for its ability to learn the relationships between words based on their context in a corpus, which is particularly useful when the relationships between words are not explicitly stated in the text. The resulting embeddings captured not only the meaning of individual words, but also their relationship with other words in the corpus. The analysis of the semantic relationships between words was performed to facilitate clustering of the data. This provided a comprehensive approach to analysing the descriptions data and enabled the identification of important patterns and relationships within the corpus.

3.2.2 Hierarchical Clustering and silhouette score

The embeddings are then clustered using the Agglomerative Clustering algorithm., which is a hierarchical clustering approach. The algorithm begins by treating each data point as an individual cluster and subsequently merges the nearest pair of clusters recursively until the desired number of clusters is attained. Next, the silhouette score is evaluated for a range of cluster values and plotted against the number of clusters (Figure 1) to determine the optimum number of clusters. The highest silhouette score denotes the optimal number of clusters. In this case, after analysing the silhouette score data, it was determined that the optimal number of clusters was 2. This implies that every time the data was sub-clustered, one cluster was bifurcated into two, and this process continued until there were only 20 data points in each cluster. This quantity is arbitrary and can be adjusted based on the data and use case. The clustering and sub-clustering produce a dendrogram (Figure 2), which is a tree-like structure that exhibits the hierarchy of clusters at various levels of similarity.

The cluster information is then stored in a new CSV file. The cluster paths can be compared to a file path, where the data points are arranged in a hierarchical structure based on their characteristics. However, instead of using names of directories and files, cluster information uses a binary representation of ones and zeros to indicate which cluster a data point belongs to at each level of subclustering.

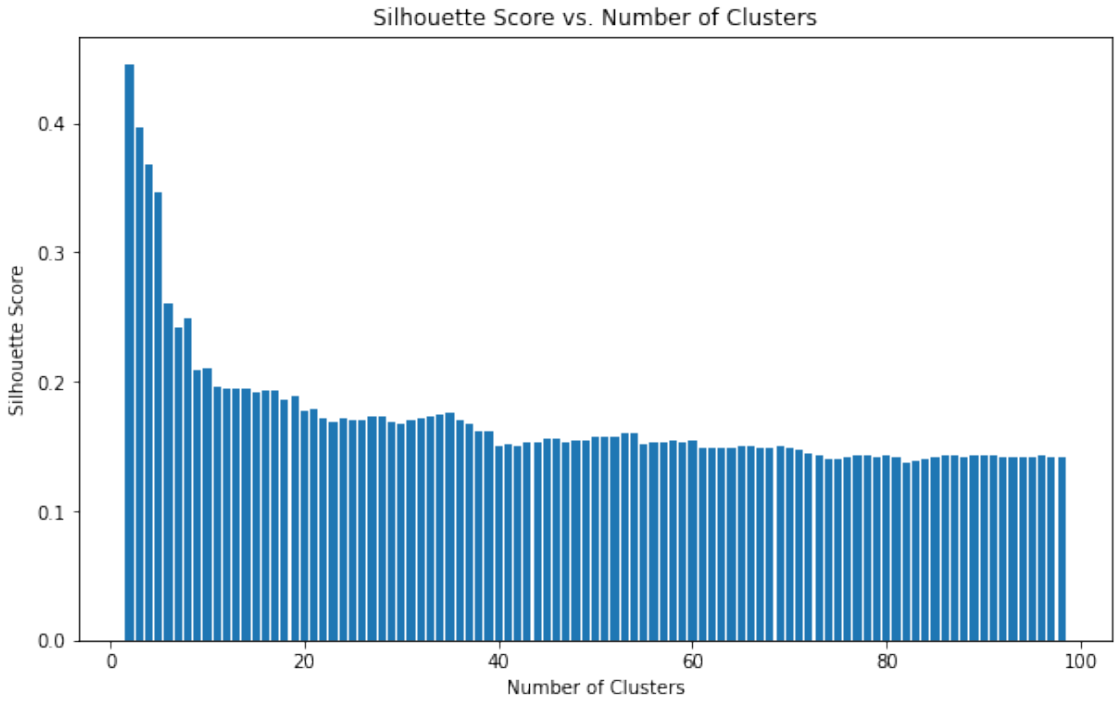

Figure 1: Silhouette Score Histogram

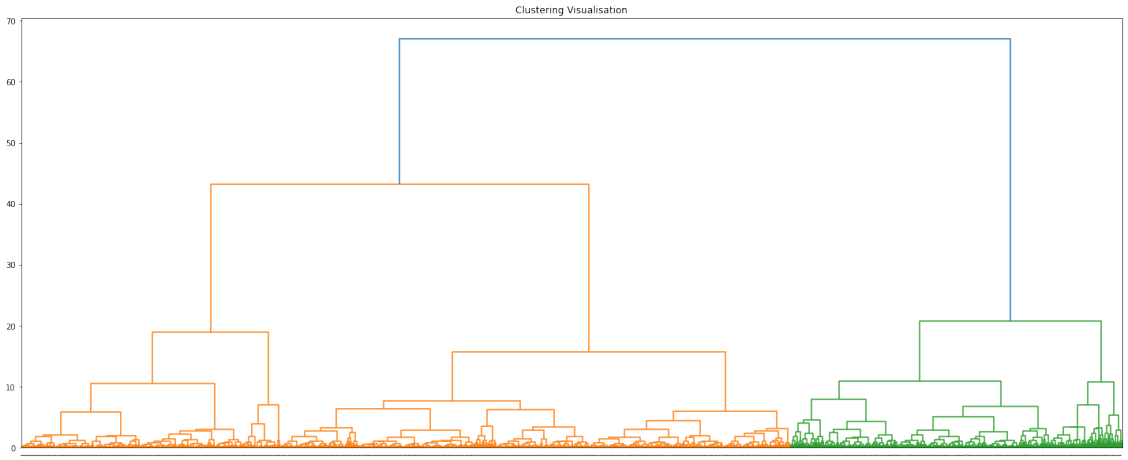

Figure 2: Dendogram

3.3 Labeling clusters using OpenAI GPT-3

Before running the clusters through the state-of-the-art GPT-3 language model, an essential preprocessing step involves merging the cluster information column with the original data set to enable the utilization of the original descriptions for the API calls. Additionally, a script is executed to compute the total count of unique clusters and sort the CSV by cluster path to facilitate the subsequent labelling process.

To accurately label each cluster, a representative sample of ten data points is randomly drawn from each unique cluster. Next, the textual descriptions of these businesses are concatenated into a unified string, which is subsequently supplied as an input to the GPT-3 API using a prompt that requires the model to discern the industry category of these businesses in three to five words.

Finally, the output from the GPT-3 API calls is preserved in a text file and aligned with the corresponding data points in the CSV file to facilitate easy retrieval of the labelling results for each cluster. The aforementioned workflow exemplifies the utilization of advanced natural language processing techniques to automate the labelling of complex data sets, thereby streamlining data analysis.

4 Results

Upon conducting this process, it was observed that certain businesses demonstrated a high level of homogeneity in their grouping, as indicated by a descriptive and precise labelling scheme. In contrast, other clusters displayed some anomalies, which resulted in a comparatively less accurate representation of the overall data set. Some clustering examples can be found in Appendix 1.

It is worth noting that this variability in the effectiveness of the clustering approach and cluster naming approach may be attributed to several factors, including the quality and comprehensiveness of the input data, the choice of clustering algorithm, and the criteria utilized for evaluating the accuracy of the resultant clusters. As such, further research and experimentation may be warranted in order to optimize the clustering process and enhance the accuracy of the cluster labelling scheme.

5 Conclusion & Future Direction

In conclusion, while the current prototype demonstrates promising potential, further refinement and development are required before it can be confidently released for widespread use. The task at hand demands a meticulous approach that prioritizes careful planning, rigorous testing, and ongoing evaluation and refinement to ultimately deliver a production-ready product, while maintaining the highest standards of quality and reliability.

- Remove company descriptions that are too short: During clustering, data points are grouped together based on their embeddings. However, if the company descriptions are too short, they may not provide enough information to effectively group the data points. Therefore, we must remove the company descriptions that are too short to ensure that the clustering algorithm has sufficient data to work with.

- Concatenate company descriptions that are too long: While short descriptions can be problematic, excessively long descriptions can also present issues for clustering algorithms. If the company descriptions are too long, they may contain irrelevant or redundant information that can confuse the algorithm. To address this, we can concatenate the lengthy descriptions into a shorter form, ensuring that they contain the most relevant information This will help to ensure that the algorithm is working with concise and informative descriptions which will improve cluster naming.

- Put company descriptions that have low affinity to a cluster in a "unclustered" cluster: When clustering, it's possible that some company descriptions may not fit well into any

of the clusters. These "outlier" entities may have characteristics that are too dissimilar to the other entities in the dataset. To account for these outliers, we can create an additional "unclustered" cluster and place any company descriptions that have low affinity to a cluster in this category. This will help us to better understand the data and identify any potential anomalies.

- Further prompt engineering to improve cluster names: Finally, to improve the quality of the cluster naming, we can further engineer the prompt used in the GPT-3 API calls to produce more accurate and descriptive cluster names.

- Visual representation of the cluster: To make the output more intuitive and accessible create a visual representation of the clusters, such as a scatter plot or a network graph, that shows the assigned cluster names. This visual output will help to provide a quick overview of the distribution of the entities and the relationships between them.

- Expansion to the full data set: The clustering algorithm is very computationally intensive, especially when dealing with large data sets. The more entities there are to cluster, the more RAM and processing power is required, which can make it difficult or even impossible to run the scripts on the full one million row data set. To address this, optimise the clustering algorithm to reduce the amount of RAM needed and improve the run time. This may involve techniques such as data sampling or parallel processing. By optimising the algorithm, we can make it feasible to run on the full data set, allowing us to extract more comprehensive and accurate insights from the data.

A Appendix 1: Clustering Examples

A.1 Name from API call: Financial Services

- Venture Crew: Venture Crew Provides consulting and investment services for investors, corporations and entrepreneurs who invest in high-tech companies.
- A3S Investments: A3S Investments is a financial service company. It is an independent investment advisory firm. It is a personalized and high-standard service to optimize resources and asset growth. It is to provide clients, according to each profile, with the most appropriate alternatives in products available on the financial market.
- Falcon Mortgage: Falcon Mortgage uses financial technology to provide mortgage services and operating information services. They provide investment products for both purchasing and financing residential properties. They offer assistance in purchasing the house. They mention their communication details on their website.

A.2 Name from API call: Advertising, software, media

- U.S. Agency for Global Media: U.S. Agency for Global Media is a media company. They provide broadcasting news services. They provide broadcasting services, voa, worldnet television and film service, and radio and tv.
- Virtual Global: Virtual Global fits digital marketing to the needs of each company. They constantly renew and adapt their processes and strategies to market changes, ensuring that the services they provide to clients are successful. They provide a comprehensive and moderate service for their website's creation, maintenance, positioning, and advertising campaigns.
- Llopart and Ramo: Llopart and Ramo is an advertising agency. It offers brand creation services, corporate identity, brand communication, product communication, art direction, illustration, and graphic design. They provide services in the field of the retail sector, catering, and food. The company provides brand communication, graphic identity on the premises, and communication of the whole range of products.